

# Covariate Selection for Generalizing Experimental Results: Application to Large-Scale Development Program in Uganda<sup>\*</sup>

Naoki Egami<sup>†</sup>

Erin Hartman<sup>‡</sup>

First Version: May 19, 2019

This Version: November 30, 2019

## Abstract

Generalizing estimates of causal effects from an experiment to a target population is of interest to scientists. However, researchers are usually constrained by available covariate information. Analysts can often collect much fewer variables from population samples than from experimental samples, which has limited applicability of existing approaches that assume rich covariate data from both experimental and population samples. In this article, we examine how to select covariates necessary for generalizing experimental results under such data constraints. In our concrete context of a large-scale development program in Uganda, although more than 40 pre-treatment covariates are available in the experiment, only 8 of them were also measured in a target population. We propose a method to estimate a *separating* set – a set of variables affecting both the sampling mechanism and treatment effect heterogeneity – and show that the population average treatment effect (PATE) can be identified by adjusting for estimated separating sets. Our algorithm only requires a rich set of covariates in the experimental data, not in the target population, by incorporating researcher-specific constraints on what variables are measured in the population data. Analyzing the development experiment in Uganda, we show that the proposed algorithm can allow for the PATE estimation in situations where conventional methods fail due to data requirements.

*Keywords:* Causal inference, External validity, Generalization, Randomized experiments

---

<sup>\*</sup>We would like to thank Christopher Blattman, Nathan Fiala, and Sebastian Martinez for sharing their data. We are also grateful to Alexander Coppock, Don Green, Chad Hazlett, Zhichao Jiang, and Soichiro Yamauchi as well as the participants of the MPSA, the UCLA IDSS workshop and the Yale Quantitative Methods Workshop, for their helpful comments on an earlier version of the paper.

<sup>†</sup>Assistant Professor (starting in 2020), Department of Political Science, Columbia University, New York NY 10027. Ph.D. Candidate, Department of Politics, Princeton University, Princeton NJ 08544. Email: negami@princeton.edu, URL: <http://scholar.princeton.edu/negami>

<sup>‡</sup>Assistant Professor of Statistics and Political Science, University of California, Los Angeles, Los Angeles, CA 90095. Email: [ekhartman@ucla.edu](mailto:ekhartman@ucla.edu), URL: [www.erinhartman.com](http://www.erinhartman.com)

# 1 Introduction

Over the last few decades, social and biomedical scientists have developed and applied an array of statistical tools to make valid causal inferences (Imbens and Rubin, 2015). In particular, randomized experiments have become the mainstay for estimating causal effects. Although many scholars agree upon the high internal validity of experimental results, there is a debate about how scientists should infer the impact of policies and interventions on broader populations (Imai *et al.*, 2008; Angrist and Pischke, 2010; Imbens, 2010; Bareinboim and Pearl, 2016; Deaton and Cartwright, 2018). This issue of generalizability (Stuart *et al.*, 2011) is pervasive in practice because randomized controlled trials are often conducted on non-representative samples (Shadish *et al.*, 2002; Druckman *et al.*, 2011; Allcott, 2015; Stuart *et al.*, 2015).

In this paper, we examine how to generalize the experimental results of the Youth Opportunities Program (YOP) in Uganda, which aims to help the poor and unemployed become self-employed artisans and increase incomes. This large scale development program, involving more than 10,000 individuals from 454 communities, was designed, implemented, and evaluated by the government of Uganda and the authors of Blattman *et al.* (2013) from 2008 to 2012. Young adults in Northern Uganda were invited to form groups and submit grant proposals for vocational training and to start independent trades. To evaluate the causal impact of the program, funding was randomly assigned among 535 eligible groups and a host of economic variables (e.g., employment and income) were measured.

The question of generalizability is especially important in this application. The aim of such development programs is elegantly noted in Duflo and Kremer (2005), “the benefits of knowing which programs work and which do not extend far beyond any program or agency, and credible impact evaluations are global public goods in the sense that they can offer reliable guidance to international organizations, governments, donors, and nongovernmental organizations (NGOs) beyond national borders.” Researchers and policy makers are not just concerned to learn about the very individuals who participated in the trial. The ultimate goal is to learn whether and how much the program can improve economic conditions in a larger target population — more than 5 million people in Northern Uganda (Blattman *et al.*, 2013).

Despite its importance, estimating population average treatment effects is not straightforward because we have to adjust for differences between experimental samples and the target

population. One pervasive question is what covariates should and can we adjust for? Although previous research shows that adjusting for a set of variables explaining sampling mechanism or treatment heterogeneity is sufficient for generalization (Stuart *et al.*, 2011; Bareinboim *et al.*, 2014), researchers are often constrained by available covariate information in applied settings.

In this paper, we address this problem of covariate selection for estimating population average treatment effects. In particular, we develop a data-driven method to estimate *a separating set* – a set of variables affecting both sampling mechanism and treatment effect heterogeneity. Recent papers show that the population average treatment effect can be identified by adjusting for this separating set (Cole and Stuart, 2010; Tipton, 2013; Pearl and Bareinboim, 2014; Kern *et al.*, 2016). In Section 3, we extend this result and show that the separating set relaxes data requirements of conventional methods by generalizing two widely-used covariate selection approaches: (1) *a sampling set* – a set of variables explaining how units are sampled into a given experiment (Pressler and Kaizar, 2013; Hartman *et al.*, 2015; Buchanan *et al.*, 2018) and (2) *a heterogeneity set*: a set of variables explaining treatment effect heterogeneity (Kern *et al.*, 2016; Nguyen *et al.*, 2017).

In Section 4, we demonstrate that such separating sets are estimable from the experimental data and provide a new estimation algorithm based on Markov random fields. This algorithm only requires that a sampling set be observed in the experimental sample, not in the target population. We estimate a separating set as a set that makes a sampling set conditionally independent of observed outcomes in the experimental data. Therefore, in contrast to conventional methods, we can exploit all covariates in the experiment to find necessary separating sets, even when there are few variables measured in both the experimental and population data.

Importantly, our proposed approach maintains a widely used assumption that a sampling set is observed in the experimental data. However, unlike many existing methods, we do not assume that a sampling set is also observed in the population data. This distinction in data requirement is subtle and yet practically essential because in many applied contexts, a larger number of covariates are measured in the experimental data than in the population data. For example, the experimental data of Blattman *et al.* (2013) contains about 40 pre-treatment covariates, even though only 8 of them are also measured in the target population. To estimate separating sets, our proposed method incorporates such user-constraints on what variables can

feasibly be collected in the target population. For instance, suppose people selected into the YOP due to social connections, which were unmeasured in the target population. Even in this scenario where conventional methods fail, the proposed method can estimate separating sets accounting for this data constraint, if any exist.

Our article builds on a growing literature on the population average treatment effect, which has two general directions. First, many previous studies have focused on articulating identification assumptions and proposing consistent estimators of the population average treatment effect (e.g., Stuart *et al.*, 2011; Hartman *et al.*, 2015; Buchanan *et al.*, 2018). In particular, recent papers explicitly show that researchers have to jointly consider treatment effect heterogeneity and the sampling mechanism (Cole and Stuart, 2010; Tipton, 2013; Pearl and Bareinboim, 2014; Kern *et al.*, 2016). These existing approaches often assume researchers have access to a large number of covariates in both the experimental sample and the non-experimental target population. In contrast, we provide a new data-driven covariate selection algorithm to find separating sets in situations where researchers have data constraints in the target population. Our focus on covariate selection is similar to recent influential work on causal directed acyclic graphs (causal DAGs) (Bareinboim *et al.*, 2014; Pearl, 2015; Bareinboim and Pearl, 2016). We differ from the DAG-based approaches in that we empirically estimate separating sets under assumptions about sampling and heterogeneity sets rather than analytically selecting separating sets from fully specified causal DAGs. Although assumptions about the entire causal DAGs are sufficient for covariate selection, the proposed algorithm can estimate separating sets under weaker assumptions about sampling and heterogeneity sets at the expense of statistical uncertainty.

Research in the second direction argues that the necessary assumptions for existing methods are often too strong in practice. Recent papers have explored methods for sensitivity analyses (Nguyen *et al.*, 2017; Andrews and Oster, 2017) and bounds (Chan, 2017) to achieve partial identification under weaker assumptions. Our paper is complementary to these approaches. We instead focus on the point identification of the population average treatment effect and alleviate strong assumptions about data requirements by adding an additional step of estimating a separating set.

## 2 Youth Opportunities Program in Uganda

As well documented by the World Bank, a large number of young adults in developing countries are unemployed or underemployed (World Bank, 2012). In addition to its direct implication to poverty, concerns for policy makers are that such large young and unemployed populations can increase risk of crime and social unrest (Blattman *et al.*, 2013). Uganda, especially conflict-affected Northern Uganda, is not an exception. According to estimates from the government, two-thirds of northern Ugandans could not meet basic needs, about 50% were illiterate, and most were underemployed in subsistence agriculture in 2006 (Government of Uganda, 2007).

In this paper, we study the Youth Opportunities Program (YOP) in Uganda, designed to help the poor and unemployed become self-employed artisans and increase incomes. This intervention is one example of widely used cash transfer programs in which participants are offered a certain amount of cash in the hope that they invest in training and start new, profitable enterprises. In 2008, the government invited young adults in Northern Uganda to form groups and submit grant proposals for how they would use a grant for vocational training and business start-up. Then, funding was randomly assigned among 535 screened, eligible applicant groups — 265 and 270 groups to treatment and control, respectively. Treatment groups received a one-time unsupervised grant worth \$7,500 on average — about \$382 per group member, roughly their average annual income. Following the original analysis, we focus on a binary treatment, whether they receive any grants or not through the YOP.

To evaluate the impact of this intervention, Blattman *et al.* (2013) surveyed 5 people per group three times over four years, resulting in a panel of 2,598 individuals after removing 79 observations due to missing data. They measured 17 outcome variables across five dimensions — employment (7), income (2), investments (3), business formality (3), and urbanization (2). They find that the effects of the YOP are large across all dimensions. Notably, after two years, the treatment groups were 4.5 times more likely to have vocational training, 2.5 times more likely to engage with a skilled trade and had 16% more hours of employment and 42% higher earnings.

Although it is unambiguous that the YOP had large, persistent positive effects on experimental subjects, it is of great policy interest to empirically investigate how much these experimental estimates are generalizable to a larger population. Estimating population average

treatment effects (PATE) can inform which specific development policies governments should scale up. While the focus of the program was on Northern Uganda as a whole, participants of the YOP were inevitably not representative, as in many other development programs. To take into account differences between experimental samples and Northern Uganda’s population, Blattman *et al.* (2013) merged their experimental samples with a 2008 population-based household survey, the Northern Uganda Survey (NUS). They adjusted for eight variables shared by experimental and population data; gender, age, urban status, marital status, school attainment, household size, durable assets, and district indicators.

In Table 1, we report estimates based on an inverse probability weighting (IPW) estimator (Stuart *et al.*, 2011) that adjusts for the original eight variables.<sup>1</sup> As a reference, we also report estimates of the average treatment effect within the experimental sample, called the sample average treatment effect (SATE). Estimates of the SATEs and PATEs have roughly the same sign, suggesting that the program will have a positive impact on a variety of outcomes even in the target population. Importantly, this finding, however, rests on an assumption that the original eight variables adjust for all relevant differences between the experimental sample and the target population. Given that the magnitude of the PATE estimates has strong implications for a cost-benefit analysis of these large-scale expensive interventions, it is critical to examine this common methodological challenge of covariate selection for generalizing experimental results.

In practice, there are several pervasive concerns about covariate selection. First, although it is common to adjust for all observed covariates shared by experimental and population data, it is unclear whether such sets of covariates include all necessary covariates for generalization. In fact, the authors carefully pay attention to this point in the original paper; “young adults are selected into our sample because of unobserved initiative, connections or affinity for entrepreneurship” (Blattman *et al.*, 2013). If there are unobserved differences between the experimental and population samples, the original PATE estimate would be biased. Second, it is also possible that the original analysis adjusted for unnecessary variables, resulting in inefficient estimators of the PATE. Miratrix *et al.* (2018) show that weighting on many variables, particularly those not highly correlated to treatment effect heterogeneity, can lead to

---

<sup>1</sup>Although the original authors rely on weighted linear regression models in their paper, we focus on the IPW estimator widely studied in the literature of generalization (e.g., Buchanan *et al.*, 2018).

	SATE estimate	Original PATE estimate		SATE estimate	Original PATE estimate
<b><u>Employment</u></b>			<b><u>Investments</u></b>		
Average employment hours	5.07 (1.09)	5.38 (3.04)	Vocational training	0.54 (0.02)	0.60 (0.05)
Agricultural	-0.05 (0.91)	-1.02 (1.91)	Hours of vocational training	348.90 (22.75)	264.21 (38.53)
Nonagricultural	5.09 (0.92)	6.68 (2.47)	Business assets	418.94 (81.71)	303.46 (65.82)
Skilled trades only	4.74 (0.62)	4.10 (1.84)	<b><u>Business Formality</u></b>		
No employment hours	-0.02 (0.01)	0.01 (0.04)	Maintain records	0.12 (0.02)	0.19 (0.06)
Any skilled trade	0.27 (0.03)	0.27 (0.06)	Registered	0.06 (0.02)	0.10 (0.06)
Works mostly in a skilled trade	0.06 (0.01)	0.04 (0.05)	Pays taxes	0.08 (0.02)	0.08 (0.05)
<b><u>Income</u></b>			<b><u>Urbanization</u></b>		
Cash earnings	13.11 (3.84)	16.85 (8.28)	Changed parish	0.06 (0.02)	0.01 (0.06)
Durable assets	0.11 (0.05)	0.16 (0.13)	Lives in Urban area	-0.01 (0.03)	0.03 (0.07)

Table 1: Estimates of Sample Average Treatment Effects and Population Average Treatment Effects based on the Original Eight Variables. *Note:* We estimated population average treatment effects (PATE) of the above 17 outcomes using an inverse probability weighting estimator with standard errors clustered by group. Weights are estimated by a logistic regression including the eight variables additively. See details of the estimation in Section 5. As a reference, we also report estimates of the sample average treatment effect (SATE).

inefficient estimation of the PATE. In this paper, we investigate necessary and sufficient sets of covariates for generalizing experimental estimates, called separating sets, and then provide a new algorithm to empirically estimate such sets. We select the separating sets under several different assumptions and assess how estimates of the PATE vary. Our reanalysis of this experiment appears in Section 5.

### 3 Separating Sets For Generalization

This section sets up the potential outcomes framework (Neyman, 1923; Rubin, 1974) for studying population average treatment effects. We review a definition of *a separating set* — a set of variables affecting both the sampling mechanism and treatment effect heterogeneity, and

then show that a sampling set and a heterogeneity set, the main focus of existing approaches, are special cases of the separating sets.

### 3.1 The Setup

We consider a scenario in which we have two data sets. Following Buchanan *et al.* (2018), we define the first sample of  $n$  individuals to be participants in a randomized experiment (“Experimental Data”) and the second data set to be a random sample of  $m$  individuals from the target population (“Population Data”). In our application, the experimental data has 2,598 individuals and the population data contains 21,348 individuals. We define a sampling indicator  $S_i$  taking 1 if unit  $i$  is in the experiment and 0 if unit  $i$  is in the target population. We assume that every unit has non-zero probability of being in the experiment. Although experimental units can be randomly sampled from the target population in ideal settings, units often non-randomly select into the experiment, as in the YOP, making the experimental sample non-representative. Note that we consider cases in which units are either in the experimental data or in the target population data, but similar results hold for cases in which the experimental sample is a subset of the target population.

Let  $T_i$  be a binary treatment assignment variable for unit  $i$  with  $T_i = 1$  for treatment and 0 for control. We define  $Y_i(t)$  to be the potential outcome variable of unit  $i$  if the unit were to receive the treatment  $t$  for  $t \in \{0, 1\}$ . In this paper, we make a stability assumption, which states that there is neither interference between units nor different versions of the treatment, either across units or settings (Rubin, 1990; Tipton, 2013; Hartman *et al.*, 2015). We define pre-treatment covariates  $\mathbf{X}_i$  to be any variables not affected by the treatment variable.

We are interested in estimating the average treatment effect in the target population. We call this causal estimand the population average treatment effect (PATE) and define it formally as follows.

**Definition 1 (Population Average Treatment Effect)**

$$\tau \equiv \mathbb{E}[Y_i(1) - Y_i(0) \mid S_i = 0],$$

where  $S_i = 0$  represents the target population data.

The treatment assignment mechanism is controlled by researchers within the experiment ( $S_i = 1$ ), but it is unknown for units in the target population ( $S_i = 0$ ; observational data). Formally, we assume that the treatment assignment is randomized within the experiment.



### Assumption 1 (Randomization in Experiment)

$$\{Y_i(1), Y_i(0), \mathbf{X}_i\} \perp\!\!\!\perp T_i \mid S_i = 1$$

This assumption holds by design in randomized experiments. Here, we consider unconditional randomization, but results in the paper can be naturally extended to settings with randomization conditional on some pre-treatment covariates. Finally, for each unit in the experimental condition, only one of the potential outcome variables can be observed, and the realized outcome variable for unit  $i$  is denoted by  $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$  (Rubin, 1974).

## 3.2 Definition of Separating Sets and Identification

Recent papers show that the PATE can be identified by a set of variables affecting both treatment effect heterogeneity and the sampling mechanism (Cole and Stuart, 2010; Tipton, 2013; Pearl and Bareinboim, 2014; Kern *et al.*, 2016). In this paper, we refer to this set as a *separating set* and investigate its statistical properties. Formally, a separating set is any set that makes the sampling indicator and treatment effect heterogeneity conditionally independent.

### Definition 2 (Separating Set)

A separating set is a set  $\mathbf{W}$  that makes the sampling indicator and treatment effect heterogeneity conditionally independent.

$$Y_i(1) - Y_i(0) \perp\!\!\!\perp S_i \mid \mathbf{W}_i. \tag{1}$$

This definition of a separating set contains two simple cases: (1) when no treatment effect heterogeneity exists and (2) when the experimental sample is randomly drawn from the target population. In both of these cases,  $\mathbf{W}_i = \{\emptyset\}$ . This separating set also encompasses two common approaches in the literature as special cases. First, researchers often employ statistical methods based on a *sampling set* – a set of all variables affecting the sampling mechanism (e.g., Stuart *et al.*, 2011). Second, researchers might adjust for a *heterogeneity set* – a set of all variables governing treatment effect heterogeneity (e.g., Kern *et al.*, 2016). Below, we formalize these sets based on the potential outcomes framework.

We define a *sampling set* as a set of variables that determines the sampling mechanism by which individuals come to be in the experimental sample. For example, when a researcher implements stratified sampling based on gender and age, the sampling set consists of those two variables. When researchers control the sampling mechanism, a sampling set is known

by design. However, when samples are selected without such an explicit sampling design, a sampling set is unknown and in practice, researchers must posit a sampling mechanism. For example, Blattman *et al.* (2013) assume that a sampling set consists of eight variables: gender, age, urban status, marital status, school attainment, household size, durable assets, and district indicators.

Formally, we can define a sampling set  $\mathbf{X}^S$  as follows.

**Definition 3 (Sampling Set)**

$$\{Y_i(1), Y_i(0), \mathbf{X}_i^{-S}\} \perp\!\!\!\perp S_i \mid \mathbf{X}_i^S \tag{2}$$

where  $\mathbf{X}^{-S}$  is a set of pre-treatment variables that are not in  $\mathbf{X}^S$ .

This conditional independence means that the sampling set is a set that sufficiently explains the sampling mechanism. Given the sampling set, the sampling indicator is independent of the joint distribution of potential outcomes and all other pre-treatment covariates. We refer to variables in the sampling set as sampling variables.

The other popular approach is to adjust for a set of all variables explaining treatment effect heterogeneity, which we call a *heterogeneity set*. Formally, we can define a heterogeneity set  $\mathbf{X}^H$  as follows.

**Definition 4 (Heterogeneity Set)**

$$Y_i(1) - Y_i(0) \perp\!\!\!\perp \{S_i, \mathbf{X}_i^{-H}\} \mid \mathbf{X}_i^H, \tag{3}$$

where  $\mathbf{X}^{-H}$  is a set of pre-treatment variables that are not in  $\mathbf{X}^H$ .

In this case, because a heterogeneity set fully accounts for treatment heterogeneity,  $Y_i(1) - Y_i(0)$  is independent of all other variables. We refer to variables in the heterogeneity set as heterogeneity variables. In our application, Blattman *et al.* (2013) discuss at least two heterogeneity variables, gender and initial credit constraints.

We want to emphasize that a sampling set and a heterogeneity set are special cases of a separating set in the sense that both sets satisfy Equation (1). Yet, there may exist many other separating sets, which we explore in Section 4.

Finally, the PATE is nonparametrically identified by adjusting for a separating set (Cole and Stuart, 2010; Tipton, 2013; Pearl and Bareinboim, 2014; Kern *et al.*, 2016).

**Result 1 (Identification of the PATE)** The PATE is identified with separating set  $\mathbf{W}_i$  under Assumption 1.

$$\tau = \int \left\{ \mathbb{E}[Y_i | T_i = 1, S_i = 1, \mathbf{W}_i = \mathbf{w}] - \mathbb{E}[Y_i | T_i = 0, S_i = 1, \mathbf{W}_i = \mathbf{w}] \right\} dF_{\mathbf{W}_i | S_i=0}(\mathbf{w}),$$

where  $F_{\mathbf{W}_i | S_i=0}(\mathbf{w})$  is the cumulative distribution function of  $\mathbf{W}$  conditional on  $S_i = 0$ .

As sampling and heterogeneity sets are special cases of a separating set, the PATE is identified with the same formula.

## 4 Identification and Estimation of Separating Sets

In this section, we first show that a variant of *separating* sets, which is sufficient for the identification of the PATE, is estimable even when a *sampling* set is unobserved in the population data as far as it is observed in the experimental data (Section 4.1). This is in contrast to existing methodologies which assume that a sampling set is observed in *both* of the experimental and population data. This distinction is subtle and yet practically important because in many applied contexts, including the YOP (Blattman *et al.*, 2013), a larger number of covariates are measured in the experimental data than in the population data. This is because, while analysts of experiments can often control what variables should be measured within the experiment, population data is usually more expensive; collected by other organizations, such as a national-level survey (the NUS in Blattman *et al.* (2013)); or otherwise impractical to collect. Thus, our focus is on this type of common research settings where analysts are able to measure more covariates in the experimental data than in the population data.

After demonstrating the identification of separating sets, we propose an algorithm to estimate separating sets using Markov random fields (Section 4.2).

### 4.1 Identification of Separating Sets

We begin with the identification of an exact separating set and then turn to the identification of a modified separating set under a weaker assumption.

First, we can estimate an exact separating set in settings where both a sampling set and a heterogeneity set are observed in the experimental data. A key feature of this result is that we only require rich covariate information about the experimental units, not the target population units, to discover separating sets, should they exist.

In many applied research contexts, however, the heterogeneity set is not readily available even in the experimental data. The fundamental problem of causal inference (Holland, 1986) states that only one of two potential outcomes are observable, which implies that the causal effect is unobserved at the unit level, and thus so is the heterogeneity set. For example, in our application, although Blattman *et al.* (2013) discuss two specific heterogeneity variables (gender and initial credit constraints), it might be unreasonable to assume away the existence of other potential heterogeneity variables.

We therefore develop an additional method to find a variant of a separating set, which we call a *marginal* separating set, using only knowledge of a sampling set, a commonly employed assumption in the extant literature. We show that a marginal separating set can be discovered when a sampling set is measured in the experimental data, but not in the target population. Although this data requirement might still be stringent in some contexts, it is much weaker than the one necessary for widely-used existing approaches based on sampling sets, which require that sampling set be measured in the population data as well as in the experimental data.

#### 4.1.1 Identification of Exact Separating Sets

We begin with settings in which a sampling set and a heterogeneity set are observed in the experimental sample. In this setting, we can use the experimental data to identify exact separating sets. Although this data requirement is still restrictive, we emphasize that it does not require rich data on the target population.

##### Setting 1 (Sampling and Heterogeneity Sets are Observed in Experiment)

Sampling set  $\mathbf{X}^S$  and heterogeneity set  $\mathbf{X}^H$  are observed in the experiment ( $S_i = 1$ ).

In this setting, a separating set is estimable as a set that makes the sampling set and the heterogeneity set conditionally independent within the experimental data.

##### Theorem 1 (Identification of Separating Sets in Experiment)

In Setting 1, for a set of pre-treatment variables  $\mathbf{W}$ , under Assumption 1,

$$\widetilde{\mathbf{X}}_i^H \perp\!\!\!\perp \widetilde{\mathbf{X}}_i^S \mid \mathbf{W}_i, T_i, S_i = 1 \implies Y_i(1) - Y_i(0) \perp\!\!\!\perp S_i \mid \mathbf{W}_i, \quad (4)$$

where  $\widetilde{\mathbf{X}}^H$  and  $\widetilde{\mathbf{X}}^S$  are the set difference  $\mathbf{X}^H \setminus \mathbf{W}$  and  $\mathbf{X}^S \setminus \mathbf{W}$ , respectively.

We provide the proof in the supplementary material. Theorem 1 states that as long as we can find a set that satisfies the testable conditional independence on the left hand side, the discovered set is guaranteed to be a separating set. That is, we can identify an exact separating set from the experimental data alone. Note that when  $\mathbf{X}^H$  and  $\mathbf{X}^S$  share some variables, those variables should always be in  $\mathbf{W}$ . Using the selected separating set, researchers can identify the PATE based on Result 1.

Intuitively, this theorem can be explained through two conceptual steps. First, because a heterogeneity set  $\mathbf{X}^H$  fully explains treatment effect heterogeneity  $Y(1) - Y(0)$ , the sampling indicator  $S$  and  $Y(1) - Y(0)$  are conditionally dependent only when  $S$  and  $\mathbf{X}^H$  are conditionally dependent. Second, because a sampling set  $\mathbf{X}^S$  fully explains the sampling indicator  $S$ ,  $S$  and  $\mathbf{X}^H$  are conditionally dependent only when  $\mathbf{X}^S$  and  $\mathbf{X}^H$  are conditionally dependent. Taken together,  $S$  and  $Y(1) - Y(0)$  are conditionally dependent only when  $\mathbf{X}^S$  and  $\mathbf{X}^H$  are conditionally dependent.

#### 4.1.2 Identification of Marginal Separating Sets

While Theorem 1 allows us to discover separating sets using the experimental data, a key challenge would be to measure both a sampling set and a heterogeneity set in the experimental data. In particular, it is often difficult to measure the heterogeneity set in practice. We show that a modified version of a separating set – a *marginal* separating set – is estimable from the experimental data under a weaker assumption. We define a marginal separating set as follows.

##### Definition 5 (Marginal Separating Set)

A marginal separating set is a set  $\mathbf{W}$  that makes the sampling indicator and the marginal distributions of potential outcomes conditionally independent.

$$Y_i(t) \perp\!\!\!\perp S_i \mid \mathbf{W}_i \quad \text{for } t = \{0, 1\}. \quad (5)$$

We refer to this as a *marginal* separating set since it renders the marginal, not the joint, distribution of potential outcomes conditionally independent of the sampling process.

Now we turn to our final setting researchers may find themselves in – that the sampling set is observed only in the experimental data. Previous work using the sampling set assumes it is measured in both the experimental sample and the target population (e.g., Cole and Stuart, 2010; Tipton, 2013; Hartman *et al.*, 2015; Buchanan *et al.*, 2018). Since researchers

often have much more control over what data is collected in the experiment, this final setting greatly relaxes the data requirements of the previous literature.

**Setting 2 (Sampling Set is Observed in Experiment)**

Sampling set  $\mathbf{X}^S$  is observed in the experimental data ( $S_i = 1$ ).

**Theorem 2 (Identification of Marginal Separating Sets in Experiment)**

In Setting 2, for a set of pre-treatment variables  $\mathbf{W}$ , under Assumption 1,

$$Y_i \perp\!\!\!\perp \mathbf{X}_i^S \mid \mathbf{W}_i, T_i, S_i = 1 \implies Y_i(t) \perp\!\!\!\perp S_i \mid \mathbf{W}_i. \tag{6}$$

We provide the proof in the supplementary material. Theorem 2 states that as long as we can find a set that makes the observed outcome  $Y$  conditionally independent of the sampling set within the experimental data, the discovered set is guaranteed to be a marginal separating set. With a large enough sample size, we can find a marginal separating set from the experimental data alone. Intuition behind this theorem is similar to the one used for Theorem 1. Because the sampling set  $\mathbf{X}^S$  fully explains the sampling indicator  $S$ , if the sampling indicator  $S$  and the potential outcome  $Y(t)$  are conditionally dependent, the sampling set  $\mathbf{X}^S$  and the observed outcome  $Y$  are also conditionally dependent. The marginal separating set may be larger than an exact separating set, as it may include prognostic covariates that explain the marginal potential outcomes but not treatment effect heterogeneity. Once we have discovered a marginal separating set using the experimental data, we can identify the PATE with this discovered set.

**Result 2 (Identification of the PATE with Marginal Separating Sets)**

When a marginal separating set  $\mathbf{W}$  is observed both in the experimental sample and the target population, the PATE is identified with the marginal separating set  $\mathbf{W}$  under Assumption 1.

$$\tau = \int \left\{ \mathbb{E}[Y_i \mid T_i = 1, S_i = 1, \mathbf{W}_i = \mathbf{w}] - \mathbb{E}[Y_i \mid T_i = 0, S_i = 1, \mathbf{W}_i = \mathbf{w}] \right\} dF_{\mathbf{W}_i|S_i=0}(\mathbf{w}).$$

We omit the proof because it is straightforward from the one of Result 1.

Finally, in Table 2, we compare existing methods with two proposed approaches. The first two rows show two common existing approaches based on sampling and heterogeneity sets, respectively. Although the identification of the PATE in those settings is straightforward, it requires rich covariate information from the target population data as well as from the

Set to Adjust For	Data Requirements	
	Experiment	Target Population
Sampling set	Sampling set	Sampling set
Heterogeneity set	Heterogeneity set	Heterogeneity set
Estimated separating set (Theorem 1 under Setting 1)	$\left\{ \begin{array}{l} \text{Sampling Set} \\ \text{Heterogeneity Set} \end{array} \right.$	User Specified Constraints
Estimated marginal separating set (Theorem 2 under Setting 2)	Sampling set	User Specified Constraints

Table 2: Identifying the PATE under different data requirements. *Note:* Many previous approaches assume that a sampling set or a heterogeneity set is measured in both the experimental sample and the target population (the first two rows). Our proposed approaches relax data requirements for the target population by introducing an additional step of estimating separating sets.

experimental sample. Our approach relaxes data requirements for the target population by introducing an additional step of estimating separating sets. In Setting 1 where we observe both a sampling set and a heterogeneity set in the experimental sample, we can identify exact separating sets from the experimental data alone (Theorem 1). Setting 2 only requires observing a sampling set in the experimental sample and we can identify marginal separating sets (Theorem 2). In the next subsection, we introduce an algorithm that can estimate separating sets subject to user specified data constraints in the target population.

## 4.2 Estimation of Separating Sets

Here, we propose an estimation algorithm to find a marginal separating set. As shown in Theorem 2, the goal is to find a set that makes a sampling set and observed outcomes conditionally independent within the experimental data. We show how to apply Markov random fields (MRFs) to encode conditional independence relationships among observed covariates and then select a separating set. A similar algorithm can be used for finding an exact separating set.

Our estimation algorithm consists of four simple steps. We provide a brief summary here

and then describe each step in order. Step 1: specify all variables in sampling set  $\mathbf{X}^S$  based on domain knowledge, some of which might not be measured in the population data. Step 2: using the experimental data alone, estimate a Markov random field over an outcome, a treatment, the sampling set and observed pre-treatment covariates. Step 3: enumerate all simple paths<sup>2</sup> from  $Y$  to  $\mathbf{X}^S$  in the estimated Markov graph. Step 4: find sets that block all the simple paths from  $Y$  to  $\mathbf{X}^S$  in the estimated Markov graph.

**Estimating Markov Random Fields.** Theorem 2 implies that we can find a marginal separating set by estimating a set of variables  $\mathbf{W}$  that satisfies the conditional independence,  $Y_i \perp\!\!\!\perp \mathbf{X}_i^S \mid \mathbf{W}_i, T_i, S = 1$ . To estimate this set, we employ a Markov random field (MRF). MRFs are statistical models that encode the conditional independence structure over random variables via graph separation rules. For example, suppose there are three random variables  $A, B$  and  $C$ . Then,  $A \perp\!\!\!\perp B \mid C$  if there is no path connecting  $A$  and  $B$  when node  $C$  is removed from the graph (i.e., node  $C$  *separates* nodes  $A$  and  $B$ ), so-called the global Markov property (Lauritzen, 1996). Using the general theory of MRFs, the estimation of a separating set can be recast as the problem of finding a set of covariates separating outcome variable  $Y$  and a sampling set  $\mathbf{X}^S$  in an estimated Markov graph. Therefore, we can find a separating set that satisfies the desired conditional independence as far as we can estimate the MRF over  $\{Y, T, \mathbf{X}^S, \mathbf{X}_0\}$  within the experimental data where we define  $\mathbf{X}_0$  to be all pre-treatment variables measured both in the experimental and population data. We define  $\mathbf{Z} \equiv \{\mathbf{X}^S, \mathbf{X}_0\}$  to be pre-treatment covariates from which we select a separating set. Note that MRFs (also known as undirected graphical models) are used here to estimate conditional independence relationships as an intermediate step of estimating separating sets; they are not used to estimate the underlying causal directed acyclic graphs (causal DAGs). In addition, we emphasize that we differ from recent causal DAGs-based approaches (Bareinboim *et al.*, 2014; Pearl and Bareinboim, 2014) in that we only rely on domain knowledge about sampling sets and we do

---

<sup>2</sup>A simple path is a path in a Markov graph that does not have repeating nodes.



not require full knowledge about underlying causal DAGs.

We use a mixed graphical model (Yang *et al.*, 2015), which allows for both continuous and categorical variables. More concretely, we assume that each node can be modeled as the exponential family distribution using the remaining variables.

$$\Pr(G_r | G_{-r}) = \exp \left\{ \alpha_r G_r + \sum_{h \neq r} \theta_{r,h} G_r G_h + \varphi(G_r) - \Phi(G_{-r}) \right\}, \quad (7)$$

where  $G_{-r}$  is a set of all random variables in a Markov graph except for variable  $G_r$ , base measure  $\varphi(G_r)$  is given by the chosen exponential family, and  $\Phi(G_{-r})$  is the normalization constant. For example, for a Bernoulli distribution, the conditional distribution can be seen as a logistic regression model.

$$\Pr(G_r | G_{-r}) = \frac{\exp(\alpha_r + \sum_{h \neq r} \theta_{r,h} G_h)}{\exp(\alpha_r + \sum_{h \neq r} \theta_{r,h} G_h) + 1}. \quad (8)$$

In general, we model each node using a generalized linear model conditional on the remaining variables. Using this setup, we can estimate the structure of the MRF by estimating parameters  $\{\theta_{r,h}\}_{h \neq r}$ ;  $\theta_{r,h} \neq 0$  for variable  $G_h$  in the neighbors of variable  $G_r$  and  $\theta_{r,h} = 0$  otherwise. We estimate each generalized linear model with  $\ell_1$  penalty to encourage sparsity (Meinshausen and Bühlmann, 2006). Finally, using the AND rule, an edge is estimated to exist between variables  $G_r$  and  $G_h$  when  $\theta_{r,h} \neq 0$  and  $\theta_{h,r} \neq 0$ . Researchers can also use an alternative OR rule (an edge exists when  $\theta_{r,h} \neq 0$  or  $\theta_{h,r} \neq 0$ ) and obtain the same theoretical guarantee of graph recovery.

**Estimating Separating Sets.** Given the estimated graphical model, we can enumerate many different separating sets. First, we focus on the estimation of a separating set of the smallest size because it often produces more stable weights and thus improves estimation accuracy. It is important to note that this separating set might not be the smallest with respect to the underlying DAG because MRFs don't encode all conditional independence relationships between variables. It is the smallest size among all separating sets estimable from MRFs.

We estimate this separating set from pre-treatment covariates  $\mathbf{Z}$  as an optimization problem. A separating set should block all simple paths between outcome  $Y$  and variables in the sampling set  $\mathbf{X}^S$ . Therefore, we first enumerate all simple paths between  $Y$  and  $\mathbf{X}^S$  and then find a minimum set of variables that intersect all paths.

Define  $q$  to denote the number of variables in  $\mathbf{Z}$ . We then define  $\mathbf{d}$  to be a  $q$ -dimensional decision vector with  $d_j$  taking 1 if we include the  $j$ th variable of  $\mathbf{Z}$  into a separating set and taking 0 otherwise. We use  $\mathbf{P}$  to store all simple paths from  $Y$  to each variable in  $\mathbf{X}^S$  where each row is a  $q$ -dimensional vector and its  $j$ th element takes 1 if the path contains the  $j$ th variable. With this setup, the estimation of the separating set of the smallest size is equivalent to the following linear programming problem given the estimated graphical model.

$$\min_{\mathbf{d}} \sum_{j=1}^q d_j \quad \text{s.t., } \mathbf{P}\mathbf{d} \geq \mathbf{1}.$$

where  $\mathbf{1}$  is a vector of ones. The constraints above ensure that all simple paths intersect with at least one variable in a selected separating set, and the objective function just counts the total number of variables to be included into a separating set. Therefore, by optimizing this problem, we can find a set of variables with the smallest size that is guaranteed to block all simple paths.

It is important to emphasize that the estimation of the Markov graph is subject to uncertainty as any other statistical methods. In our application, we incorporate uncertainties about set estimation through bootstrap. We also investigate accuracy of the proposed algorithm through simulation studies in the supplementary material. We find that estimators based on estimated separating sets often have similar standard errors to the ones based on the true sampling set. Although our approach introduces an additional estimation step of finding separating sets to relax data requirements, it does not suffer from substantial efficiency loss.

**Incorporating Users' Constraints.** One advantage of our approach is that we can allow the flexibility for researchers to explicitly specify variables that they cannot measure in the target population. This is important in practice because it is often the case that researchers can

measure a large number of covariates in the experimental data but they can collect relatively few variables in the target population. We can easily adjust the previous optimization problem to account for this restriction. Define  $\mathbf{u}$  to be a  $q$ -dimensional vector with  $u_j$  taking 1 if we want to exclude the  $j$  th variable of  $\mathbf{Z}$  from a separating set and taking 0 otherwise. As we define  $\mathbf{X}_0$  to be those variables observed in both the experimental sample and the target population,  $\mathbf{u}$  will place constraints on those covariates in  $\mathbf{X}^S$  that are unobservable. Then, the optimization problem above changes as follows.

$$\min_{\mathbf{d}} \sum_{j=1}^q d_j \quad \text{s.t., } \mathbf{P}\mathbf{d} \geq \mathbf{1} \quad \text{and} \quad \mathbf{u}^\top \mathbf{d} = 0$$

In practice, it is possible that there exists no separating set, subject to user constraints. In our example, a true separating set could include social connections, which are not measured in the Northern Uganda Survey (the population data). In this case, there is no feasible separating set and our algorithm finds no separating set.

### 4.3 Estimation of Population Average Treatment Effect

To estimate the PATE with estimated separating sets, we use an inverse probability weighting estimator. First, we estimate a probability of being in the experiment  $\Pr(S_i = 1 \mid \mathbf{W}_i)$ , for example, using a logistic regression (Stuart *et al.*, 2011; Westreich *et al.*, 2017). Following Buchanan *et al.* (2018), we stack the experimental data and the population data, and  $S_i = 1$  ( $S_i = 0$ ) indicates that unit  $i$  belongs to the experimental data (the population data). We can then estimate weights as

$$\pi_i = \frac{1}{\Pr(S_i = 1 \mid \mathbf{W}_i)} \times \frac{\Pr(S_i = 0 \mid \mathbf{W}_i)}{\Pr(S_i = 0)}, \quad (9)$$

where a usual inverse probability is adjusted by  $\Pr(S_i = 0 \mid \mathbf{W}_i) / \Pr(S_i = 0)$  because the PATE is defined only with the population data, i.e.,  $\mathbb{E}[Y_i(1) - Y_i(0) \mid S_i = 0]$ . Finally, we compute the inverse probability weighting estimator (Stuart *et al.*, 2011).

$$\hat{\tau} \equiv \frac{\sum_{i; S_i=1} \pi_i p_i T_i Y_i}{\sum_{i; S_i=1} \pi_i p_i T_i} - \frac{\sum_{i; S_i=1} \pi_i (1 - p_i) (1 - T_i) Y_i}{\sum_{i; S_i=1} \pi_i (1 - p_i) (1 - T_i)}, \quad (10)$$

where  $p_i \equiv \Pr(T_i = 1 \mid S_i = 1, \mathbf{W}_i)$  is known by the experimental design. We prove its consistency in the supplementary material.

## 5 Empirical Analysis

Applying the proposed method, we examine the YOP described in Section 2. Our focus is on a central methodological challenge of covariate selection. In the original analysis, the authors adjusted for all eight variables shared by the experimental and population data. However, as noted in the original paper, it is unknown whether the original eight variables is a separating set necessary for estimating PATEs. To tackle this pervasive concern, we employ the proposed approach and select a separating set under two different assumptions about a sampling set and a heterogeneity set.

First, we incorporate domain knowledge about a heterogeneity set, while we maintain the original assumption about a sampling set. As explained in Section 3, by combining substantive information about a sampling set and a heterogeneity set, we can find a separating set, which can be much smaller than each one of the two. Relying on this smaller separating set, we find that point estimates are similar to estimates based on the original sampling set, but standard errors of the proposed approach are much smaller for 16 out of 17 outcomes that the original analysis studied. Incorporating domain knowledge about a heterogeneity set can help us find a smaller set of variables sufficient for the PATE estimation, thereby improving efficiency.

Second, we relax the original assumption about a sampling set — the shared eight variables contain all relevant variables, and we allow for two additional unobserved variables. In the conventional approach based on a sampling set, researchers cannot estimate PATEs under this assumption. In contrast, the proposed approach estimated appropriate separating sets for 12 out of 17 outcomes and we find that PATE estimates for those outcomes are robust to the two additional unobserved sampling variables. At the same time, we reveal that estimated PATEs are sensitive to the original assumption about the sampling mechanism for the other

5 outcomes.

## 5.1 Incorporating Domain Knowledge on Heterogeneity Set

To begin with, we maintain an assumption about a sampling set in the original analysis, i.e.,  $\mathbf{X}^S = \{\text{Gender, Age, Urban, Marital status, School attainment, Household size, Durable assets, District}\}$ . Although the original analysis relies only on this knowledge of the sampling set for the PATE estimation, the authors also carefully discuss a heterogeneity set in their paper. In particular, they discuss two variables: gender and initial credit constraints. There are two natural covariates in the experimental data that capture these concepts, `Gender` and `Initial Saving`, respectively. Importantly, however, `Initial Saving` is measured only in the experimental sample and not in the target population data. Thus,  $\mathbf{X}^H = \{\text{Gender, } \boxed{\text{Initial Saving}}\}$  where the square box represents a variable unmeasured in the target population.

In existing approaches, when a subset of heterogeneity variables are unmeasured in the target population as in this case, it is difficult to incorporate such domain knowledge into analysis and researchers often ignore the heterogeneity set altogether. In contrast, our proposed method uses knowledge about the heterogeneity set to estimate an exact separating set, which is potentially smaller than the observed sampling set and thus can increase the estimation accuracy for the PATE estimation. We first estimate a Markov random field over the union of sampling and heterogeneity sets within the experimental data. Then, we select an exact separating set that makes sampling set  $\mathbf{X}^S$  and heterogeneity set  $\mathbf{X}^H$  conditionally independent under a constraint that `Initial Saving` is unmeasured in the population data and cannot be selected. To take into account uncertainties, we estimate Markov random fields and select exact separating sets in each of 1000 bootstrap samples.

Figure 1 reports the results. The left panel (a) shows the proportion of each variable being estimated to be in an exact separating set over 1000 bootstrap samples. As the definition of separating sets (Definition 2) implies, the intersection of sampling and heterogeneity set,

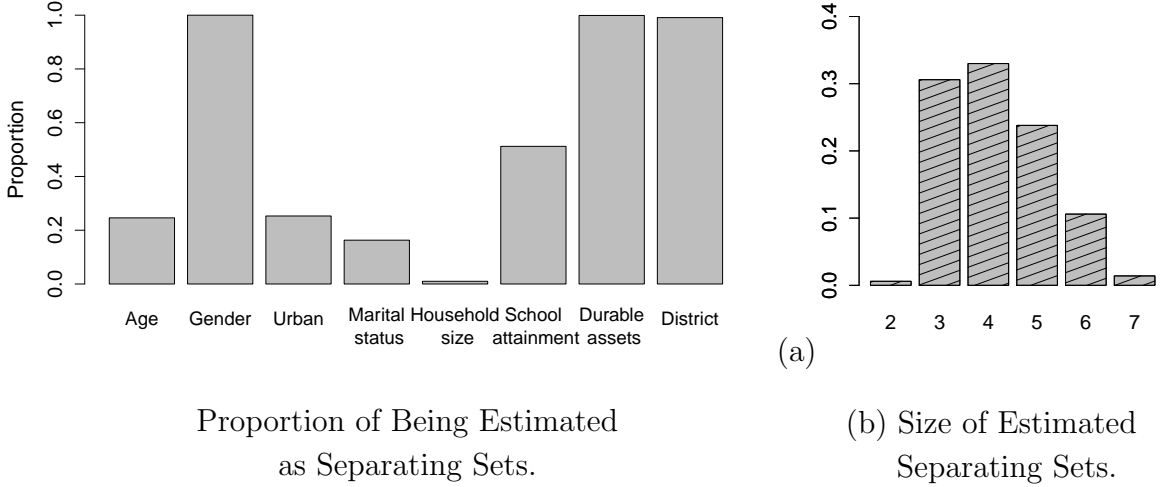


Figure 1: Estimated Exact Separating Sets. *Note:* Panel (a) shows the proportion of each variable being estimated to be in an exact separating set over 1000 bootstrap samples. Panel (b) reports the size of estimated exact separating sets over 1000 bootstrap samples.

`Gender`, is always selected. In addition, `Durable assets` and `District` are selected almost always. Importantly, when we look at the size of estimated exact separating sets (the right panel (b)), it is often much smaller than the original size of eight (the mean size is 4.17). This means that even when a sampling set is sufficient for estimating the PATE, researchers can find a smaller separating set by incorporating domain knowledge of heterogeneity sets with the proposed approach.

If assumptions about the sampling set and the heterogeneity set hold, estimators based on the original sampling set  $\mathbf{X}^S$  and on the estimated separating sets  $\mathbf{W}$  are both consistent. However, standard errors of the latter might be smaller because corresponding estimated weights might be more stable.

To estimate the PATEs, we use the inverse probability weighting estimator proposed in Section 4.3. First, we estimate weights using the following logistic regression.

$$\text{logit}\{\Pr(S_i = 1 \mid \mathbf{C}_i)\} = \alpha_0 + \mathbf{C}_i^\top \beta, \quad (11)$$

where  $\mathbf{C} = \mathbf{X}^S$  for the estimator based on the original sampling set and  $\mathbf{C} = \mathbf{W}$  for our proposed estimator. We stack the experimental data (sample size = 2,598) and the

population data (sample size = 21,348) and  $S_i = 1$  ( $S_i = 0$ ) indicates that unit  $i$  belongs to the experimental data (the population data). We can then estimate weights as  $\hat{\pi}_i = 1/\widehat{\Pr}(S_i = 1 | \mathbf{C}_i) \times \widehat{\Pr}(S_i = 0 | \mathbf{C}_i)/\widehat{\Pr}(S_i = 0)$ , as proposed in Section 4.3. Note that treatment assignment probability in the experiment  $\Pr(T_i = 1 | S_i = 1, \mathbf{W}_i)$  is equal to  $\Pr(T_i = 1 | S_i = 1, \mathbf{D}_i)$  where  $\mathbf{D}_i$  is a vector indicating 14 districts, because the treatment randomization was stratified by districts (Blattman *et al.*, 2013). Standard errors are clustered by group as done in the original analysis. Note that the difference between the estimator based on the original sampling set and our proposed estimator comes only from the selection of covariates  $\mathbf{C}$  in the estimation of weights.

We report results in Table 3. Effects of the YOP are large and positive across many outcomes even among the broader target population. For example, the average employment hours would increase by 4.62 hours (19% increase compared to the control group), monthly cash earnings would increase by 12,490 Uganda shilling (37% increase), and a proportion of people enrolled in vocational training would increase by 54 percentage points (360% increase). Comparing estimates based on the original sampling set and those based on the proposed separating set, we reveal that point estimates are often smaller than estimates with the original eight variables, although differences between them are not statistically significant at the conventional 0.05 level. This is expected because both estimators are consistent under the assumption that both specified sampling and heterogeneity sets are correct. More interestingly, we find that, for 16 out of 17 outcomes, standard errors of estimators based on the estimated separating sets are smaller than those based on the original sampling set. On average, standard errors of the proposed approach are about 34% smaller. For the outcome “Works mostly in a skilled trade,” the standard error reduces more than 60%. This shows that by incorporating domain knowledge about heterogeneity sets, we can estimate smaller separating sets, which often improve efficiency.

	Original estimate	Sep. Set estimate		Original estimate	Sep. Set estimate
<b><u>Employment</u></b>			<b><u>Investments</u></b>		
Average employment hours	5.38 (3.04)	4.62 (1.64)	Vocational training	0.60 (0.05)	0.54 (0.04)
Agricultural	-1.02 (1.91)	0.02 (1.27)	Hours of vocational training	264.21 (38.53)	327.6 (30.42)
Nonagricultural	6.68 (2.47)	4.60 (1.32)	Business assets	303.46 (65.82)	411.52 (115.07)
Skilled trades only	4.10 (1.84)	4.39 (0.82)	<b><u>Business Formality</u></b>		
No employment hours	0.01 (0.04)	-0.03 (0.02)	Maintain records	0.19 (0.06)	0.17 (0.04)
Any skilled trade	0.27 (0.06)	0.27 (0.04)	Registered	0.10 (0.06)	0.07 (0.03)
Works mostly in a skilled trade	0.04 (0.05)	0.04 (0.02)	Pays taxes	0.08 (0.05)	0.07 (0.03)
<b><u>Income</u></b>			<b><u>Urbanization</u></b>		
Cash earnings	16.85 (8.28)	12.49 (4.25)	Changed parish	0.01 (0.06)	0.03 (0.04)
Durable assets	0.16 (0.13)	0.15 (0.09)	Lives in Urban area	0.03 (0.07)	0.00 (0.03)

Table 3: Estimates of Population Average Treatment Effects based on the Original Set and the Estimated Separating Set. *Note:* We estimated population average treatment effects of 17 outcomes using weights based on the original eight variables (“Original estimate”) and the estimated exact separating set (“Sep. Set estimate”). Standard errors of the proposed estimators are smaller for 16 out of 17 outcomes.

## 5.2 Accounting for Unobserved Sampling Set

In the previous analysis, we maintained the original authors’ assumption about the sampling set and additionally take into account the assumption about the heterogeneity set. Here, we focus on estimating PATEs under weaker assumptions and directly address a concern noted in the original paper that the shared eight variable might not contain all relevant variables. In particular, Blattman *et al.* (2013) discuss two potentially problematic variables. First, the authors are concerned that when the government screened applications at



the village level, people with more social connections may have received some privilege. Second, people with “affinity for entrepreneurship” (Blattman *et al.*, 2013) might have been more likely to apply for the program in the first place. To account for these two sources of sample selection, we assume that a true sampling set contains two additional variables: (1) **Connection**, the number of community groups that a respondent belongs to, as a measure of social connections, and (2) **Business Advice**, total hours spent getting business advice in last 7 days, as a measure of initial motivation and affinity for entrepreneurship. Importantly, both of these two variables are not measured in the population data. Therefore,  $\mathbf{X}^S = \{\text{Gender, Age, Urban, Marital status, School attainment, Household size, Durable assets, District, Connection, Business Advice}\}$  where the last two variables are measured only in the experiment and not in the population data. Moreover, we don’t make any assumption about heterogeneity sets. Under this assumption, the current practice based on sampling sets or heterogeneity sets cannot estimate any PATEs; weights can be estimated only when sampling sets or heterogeneity sets are measured in both the experimental and population data. In contrast, the proposed method can select appropriate separating sets, should they exist, under such data constraints.

There are two questions of interest for each outcome; (1) Can we find a separating set and estimate the PATE? (2) If we can estimate the PATE, is an estimate different from the one based on the original eight variables? We estimate marginal separating sets using the proposed algorithm. For each outcome  $Y$ , we first estimate a Markov random field and then select a separating set that makes outcome  $Y$  and sampling set  $\mathbf{X}^S$  conditionally independent under a constraint that the two unobserved variables (**Connection**, **Business Advice**) cannot be selected. When the algorithm can find no separating set under the constraint, we call it an “infeasible solution.” To take into account uncertainties over this covariate selection, we estimate Markov random fields and select separating sets in each of 1000 bootstrap samples.

We begin by computing proportions of infeasible solutions among the 1000 bootstraps

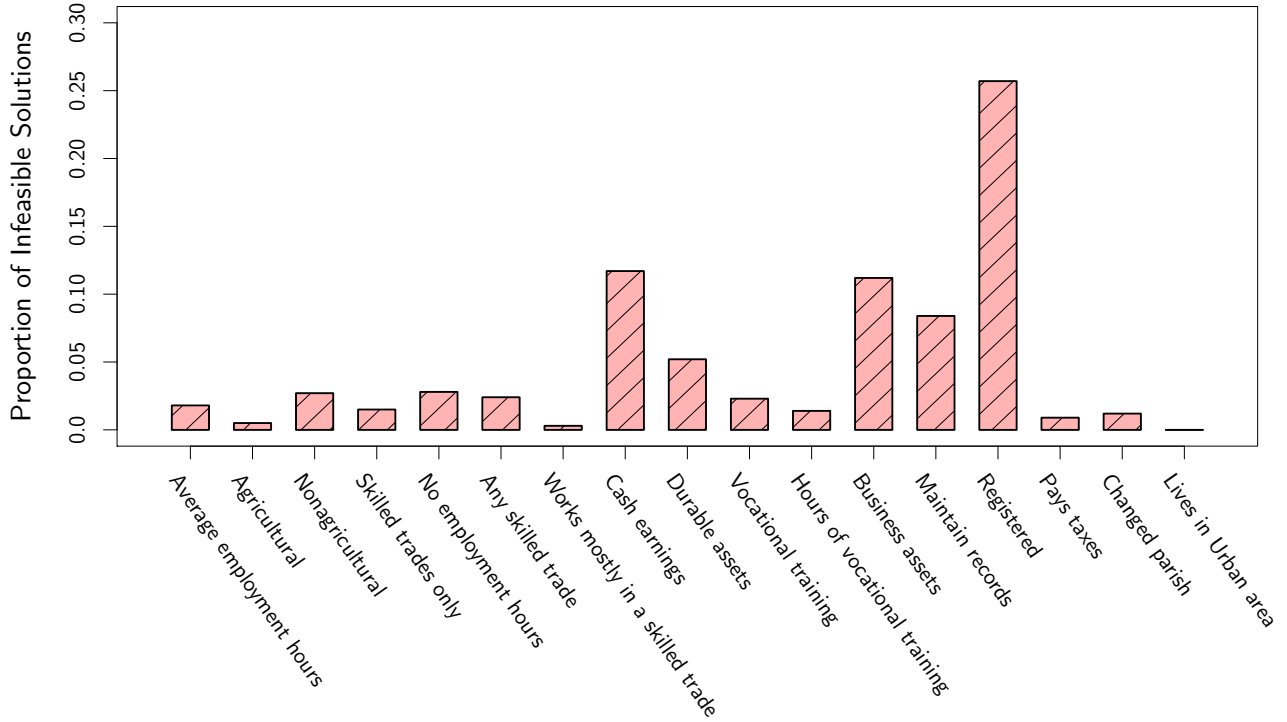


Figure 2: Proportions of Infeasible Solutions. *Note:* For 17 outcomes, we estimated marginal separating sets under a constraint that two sampling variables are unobserved in the population data. The figure shows the proportion of infeasible solution for each outcome.

(Figure 2). Proportions vary across outcomes, ranging from 0.0% (“Lives in Urban area”) to 25.7% (“Registered”), and on average, 4.71%. Given that the current practice just based on sampling or heterogeneity sets cannot estimate PATEs for any outcomes, it is interesting that the proportions of infeasible solutions are smaller than 5% for 12 out of 17 outcomes. For the remaining five outcomes, the average proportion of infeasible solutions is 12.4%, suggesting that the PATE estimates for these outcomes are sensitive to unobserved sampling variables, **Connection and Business Advice**.

For four outcomes that have less than 1% of infeasible solutions,<sup>3</sup> we also report estimates with 95% confidence intervals in Figure 3. We find that point estimates are similar to the original estimates and yet, standard errors are smaller than the original ones. That is, estimates of the PATEs are robust to alternative separating sets, i.e., even if the sampling set includes

<sup>3</sup>“Employment in agriculture” [0.5%], “Works mostly in a skilled trade” [0.3%], “Lives in Urban area” [0.0%], “Pays taxes” [0.9%].

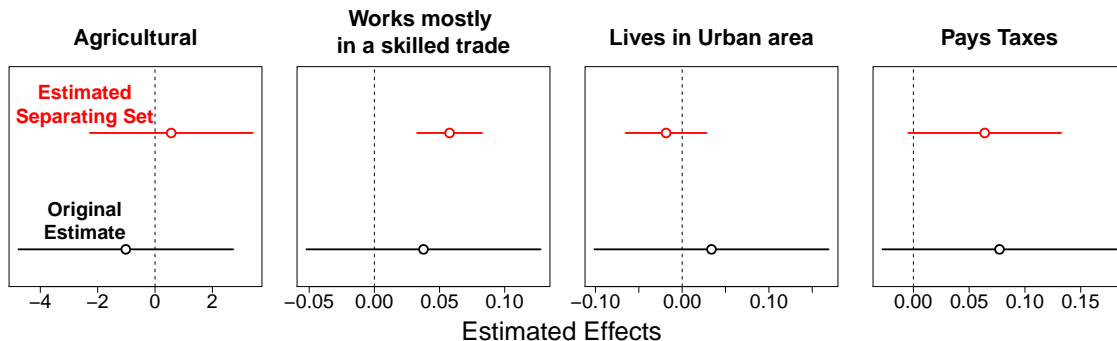


Figure 3: Estimates of Population Average Treatment Effects based on the Original Sets and Estimated Marginal Separating Sets. *Note:* We estimated population average treatment effects for 4 outcomes that have estimated proportions of infeasible solutions below 1%. Weights are based on the original eight variables (“Original”) and estimated marginal separating sets (“Estimated Separating Set”).

additional unobserved variables, substantive conclusions are similar. This result demonstrates that the proposed algorithm of selecting separating sets allows researchers to estimate PATEs in situations where previous methods could not.

## 6 Concluding Remarks

The increased emphasis on well-identified causal effects in the social and biomedical sciences can sometimes lead researchers to narrow the focus of their research question and limit their findings to the experimental sample. However, primary research questions are often driven by the need to discover the impact of an intervention on a broader population. The extant literature has focused on the mathematical underpinnings concerning the generalizability of experimental evidence. The aim of this paper is to provide applied researchers with a means for uncovering a separating set using the experimental data alone.

Building on previous approaches to generalization, we clarify the role of the separating set – and its relationship to the sampling mechanism and treatment effect heterogeneity – in identification of population average treatment effects. This framework makes clear that there are many possible covariate sets researchers can use for the recovery of population effects, and it allows us to develop a new algorithm that can incorporate researchers’ data constraints on

the target population.

As a concrete context, we focus on the YOP in Uganda, designed to help the poor and unemployed become self-employed artisans and increase their incomes. For these types of large-scale development programs, potential benefits and necessity of generalization are well known among researchers and policy makers. However, analysts are often constrained by available covariate information, which limits applicability of existing approaches that assume rich covariate data from both the experimental and population samples. Our proposed algorithm can help researchers to estimate appropriate separating sets, if any should exist, even under such data constraints. We find that by incorporating domain knowledge about heterogeneity sets, which is often overlooked in the PATE estimation, we can substantially improve efficiency. We also reveal that the proposed algorithm can find separating sets for 12 out of 17 outcomes, even if we allow for two additional sampling variables that are not measured in the population.

Identifying population effects remains a challenging task for experimental researchers. The results here suggest researchers can increase a chance of generalization by collecting rich covariate information on their experimental subjects, even when their capacity of the population data collection is limited.

## References

- Allcott, H. (2015). Site selection bias in program evaluation. *Quarterly Journal of Economics*, pages 1117–1165.
- Andrews, I. and Oster, E. (2017). Weighting for External Validity. *NBER working paper*.
- Angrist, J. D. and Pischke, J.-S. (2010). The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics. *Journal of Economic Perspectives*, **24**(2), 3–30.
- Bareinboim, E. and Pearl, J. (2016). Causal Inference and the Data-Fusion Problem. *Proceedings of the National Academy of Sciences*, **113**(27), 7345–7352.

- Bareinboim, E., Tian, J., and Pearl, J. (2014). Recovering from Selection Bias in Causal and Statistical Inference. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.
- Blattman, C., Fiala, N., and Martinez, S. (2013). Generating Skilled Self-Employment in Developing Countries: Experimental Evidence from Uganda. *The Quarterly Journal of Economics*, **129**(2), 697–752.
- Buchanan, A. L., Hudgens, M. G., Cole, S. R., Mollan, K. R., Sax, P. E., Daar, E. S., Adimora, A. A., Eron, J. J., and Mugavero, M. J. (2018). Generalizing evidence from randomized trials using inverse probability of sampling weights. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **95**, 1082.
- Chan, W. (2017). Partially Identified Treatment Effects for Generalizability. *Journal of Research on Educational Effectiveness*, **10**(3), 646–669.
- Cole, S. R. and Stuart, E. A. (2010). Generalizing Evidence From Randomized Clinical Trials to Target Populations: The ACTG 320 Trial. *American journal of epidemiology*, **172**(1), 107–115.
- Deaton, A. and Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*.
- Druckman, J. N., Green, D. P., Kuklinski, J. H., and Lupia, A. (2011). *Cambridge Handbook of Experimental Political Science*. Cambridge University Press.
- Dufo, E. and Kremer, M. (2005). Use of Randomization in the Evaluation of Development Effectiveness. In G. Pitman, O. Feinstein, and G. Ingram, editors, *Evaluating Development Effectiveness*, pages 205–232. Transaction Publishers, New Brunswick, NJ.
- Government of Uganda (2007). National Peace, Recovery and Development Plan for Northern Uganda: 2006–2009. Technical report, Government of Uganda, Kampala.
- Hartman, E., Grieve, R., Ramsahai, R., and Sekhon, J. S. (2015). From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **178**(3), 757–778.
- Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, **81**(396), 945–960.
- Imai, K., King, G., and Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **171**(2), 481–502.

- Imbens, G. W. (2010). Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009). *Journal of Economic Literature*, **48**(2), 399–423.
- Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.
- Kern, H. L., Stuart, E. A., Hill, J., and Green, D. P. (2016). Assessing Methods for Generalizing Experimental Impact Estimates to Target Populations. *Journal of Research on Educational Effectiveness*, **9**(1), 103–127.
- Lauritzen, S. L. (1996). *Graphical Models*. Clarendon Press, Oxford.
- Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, **34**(3), 1436–1462.
- Miratrix, L. W., Sekhon, J. S., Theodoridis, A. G., and Campos, L. F. (2018). Worth weighting? how to think about and use weights in survey experiments. *Political Analysis*, **26**(3), 275–291.
- Neyman, J. (1923). On the Application of Probability Theory to Agricultural Experiments. Essay on Principles (with discussion). Section 9 (translated). *Statistical Science*, **5**(4), 465–472.
- Nguyen, T. Q., Ebnesajjad, C., Cole, S. R., and Stuart, E. A. (2017). Sensitivity analysis for an unobserved moderator in RCT-to-target-population generalization of treatment effects. *The Annals of Applied Statistics*, **11**(1), 225–247.
- Pearl, J. (2000). *Causality: Models, Reasoning and Inference*. Cambridge University Press.
- Pearl, J. (2015). Generalizing Experimental Findings. *Journal of Causal Inference*, **3**(2), 259–266.
- Pearl, J. and Bareinboim, E. (2014). External Validity: From Do-Calculus to Transportability Across Populations. *Statistical Science*, **29**(4), 579–595.
- Pressler, T. R. and Kaizar, E. E. (2013). The use of propensity scores and observational data to estimate randomized controlled trial generalizability bias. *Statistics in medicine*, **32**(20), 3552–3568.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, **66**(5), 688.
- Rubin, D. B. (1990). Comment on J. Neyman and causal inference in experiments and observational studies: “On the application of probability theory to agricultural experiments. Essay on principles. Section 9” [Ann. Agric. Sci. 10 (1923), 1–51]. *Statistical Science*, **5**(4), 472–480.

- Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.
- Stuart, E. A., Cole, S. R., Bradshaw, C. P., and Leaf, P. J. (2011). The Use of Propensity Scores to Assess the Generalizability of Results from Randomized Trials. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **174**(2), 369–386.
- Stuart, E. A., Bradshaw, C. P., and Leaf, P. J. (2015). Assessing the Generalizability of Randomized Trial Results to Target Populations. *Prevention Science*, **16**(3), 475–485.
- Tipton, E. (2013). Improving Generalizations From Experiments Using Propensity Score Subclassification: Assumptions, Properties, and Contexts. *Journal of Educational and Behavioral Statistics*, **38**(3), 239–266.
- Westreich, D., Edwards, J. K., Lesko, C. R., Stuart, E., and Cole, S. R. (2017). Transportability of Trial Results Using Inverse Odds of Sampling Weights. *American Journal of Epidemiology*, **186**(8), 1010–1014.
- World Bank (2012). World development report 2013: Jobs. Technical report, World Bank, Washington, DC.
- Yang, E., Ravikumar, P., Allen, G. I., and Liu, Z. (2015). Graphical models via univariate exponential family distributions. *Journal of Machine Learning Research*, **16**(1), 3813–3847.

# Supplementary Material

## Covariate Selection for Generalizing Experimental Results

### SI-1 Proof of Theorems

Here, we provide proofs for the theorems presented in the paper.

#### SI-1.1 Proof of Theorem 1

In this proof, we assume that the separating set  $\mathbf{W}$  is disjoint with the sampling set  $\mathbf{X}^S$  and the heterogeneity set  $\mathbf{X}^H$  for simpler notations. The same proof applies to the case in which some variables of the sampling set or the heterogeneity set are in the separating set. First, we have

$$\mathbf{X}^H \perp\!\!\!\perp \mathbf{X}^S \mid \mathbf{W}, T, S = 1. \quad (12)$$

From Random Treatment Assignment (Assumption 1), we have

$$T \perp\!\!\!\perp \mathbf{X}^S \mid \mathbf{W}, S = 1. \quad (13)$$

Combining equations (12) and (13) (Contraction in Pearl (2000)),

$$\{\mathbf{X}^H, T\} \perp\!\!\!\perp \mathbf{X}^S \mid \mathbf{W}, S = 1,$$

which implies  $\mathbf{X}^H \perp\!\!\!\perp \mathbf{X}^S \mid \mathbf{W}, S = 1$ . Given that the conditional independence structure of  $(\mathbf{X}^H, \mathbf{X}^S, \mathbf{W})$  is the same under  $S = 1$  and  $S = 0$  (because  $S$  only changes the treatment assignment), we have

$$\mathbf{X}^H \perp\!\!\!\perp \mathbf{X}^S \mid \mathbf{W}, S. \quad (14)$$

From the definition of the sampling variable,

$$\mathbf{X}^H \perp\!\!\!\perp S \mid \mathbf{W}, \mathbf{X}^S. \quad (15)$$

Combining equations (14) and (15) (Intersection (Pearl, 2000)), we have

$$\mathbf{X}^H \perp\!\!\!\perp \{S, \mathbf{X}^S\} \mid \mathbf{W},$$

which implies

$$\mathbf{X}^H \perp\!\!\!\perp S \mid \mathbf{W}. \quad (16)$$



Additionally, based on the definition of the heterogeneity set,

$$Y(1) - Y(0) \perp\!\!\!\perp S \mid \mathbf{W}, \mathbf{X}^H. \quad (17)$$

Therefore, by combining equations (16) and (17) based on Contraction in Pearl (2000),

$$\{Y(1) - Y(0), \mathbf{X}^H\} \perp\!\!\!\perp S \mid \mathbf{W},$$

which implies  $Y(1) - Y(0) \perp\!\!\!\perp S \mid \mathbf{W}$ . □

## SI-1.2 Proof of Theorem 2

First, we have

$$Y \perp\!\!\!\perp \mathbf{X}^S \mid \mathbf{W}, T, S = 1. \quad (18)$$

From Random Treatment Assignment(Assumption 1), we have

$$T \perp\!\!\!\perp \mathbf{X}^S \mid \mathbf{W}, S = 1. \quad (19)$$

Combining equations (18) and (19) (Contraction in Pearl (2000)),

$$\{Y, T\} \perp\!\!\!\perp \mathbf{X}^S \mid \mathbf{W}, S = 1,$$

which implies

$$Y(t) \perp\!\!\!\perp \mathbf{X}^S \mid \mathbf{W}, S = 1. \quad (20)$$

Given that the conditional independence structure of  $(Y(1), Y(0), \mathbf{X}^S, \mathbf{W})$  is the same under  $S = 1$  and  $S = 0$  (because  $S$  only changes the treatment assignment, relationship for potential outcomes and pre-treatment variables would not change), we have

$$Y(t) \perp\!\!\!\perp \mathbf{X}^S \mid \mathbf{W}, S, \quad (21)$$

for  $t = \{0, 1\}$ .

From the definition of the sampling variable, for  $t = \{0, 1\}$ ,

$$Y(t) \perp\!\!\!\perp S \mid \mathbf{W}, \mathbf{X}^S. \quad (22)$$

Combining equations (21) and (22) (Intersection in Pearl (2000)), we have

$$Y(t) \perp\!\!\!\perp \{S, \mathbf{X}^S\} \mid \mathbf{W},$$

which implies

$$Y(t) \perp\!\!\!\perp S \mid \mathbf{W}$$

for  $t = \{0, 1\}$ . This completes the proof. □

## SI-2 IPW Estimator

Here, we show that  $\hat{\tau} \xrightarrow{P} \mathbb{E}[Y_i(1) - Y_i(0) \mid S_i = 0]$ .

**Proof** First, we rewrite the IPW estimator as follows.

$$\hat{\tau} = \frac{\frac{1}{n+m} \sum_i S_i \pi_i p_i T_i Y_i}{\frac{1}{n+m} \sum_i S_i \pi_i p_i T_i} - \frac{\frac{1}{n+m} \sum_i S_i \pi_i (1-p_i)(1-T_i) Y_i}{\frac{1}{n+m} \sum_i S_i \pi_i (1-p_i)(1-T_i)}, \quad (23)$$

where  $n$  ( $m$ ) is the sample size of the experimental data (the population data). By the law of large number,

$$\begin{aligned} \frac{1}{n+m} \sum_i S_i \pi_i p_i T_i &\xrightarrow{P} \mathbb{E}[S_i \pi_i p_i T_i] = \mathbb{E}_{\mathbf{W}} \{ \pi_i \Pr(S_i = 1 \mid \mathbf{W}_i) p_i \Pr(T_i = 1 \mid S_i = 1, \mathbf{W}_i) \} \\ &= \mathbb{E}_{\mathbf{W}} \left\{ \frac{\Pr(S_i = 0 \mid \mathbf{W}_i)}{\Pr(S_i = 0)} \right\} = 1. \end{aligned}$$

Similarly,  $\frac{1}{n+m} \sum_i S_i \pi_i (1-p_i)(1-T_i) \xrightarrow{P} 1$ . Again, by the law of large number,

$$\frac{1}{n+m} \sum_i S_i \pi_i p_i T_i Y_i \xrightarrow{P} \mathbb{E}[S_i \pi_i p_i T_i Y_i], \quad \frac{1}{n+m} \sum_i S_i \pi_i (1-p_i)(1-T_i) Y_i \xrightarrow{P} \mathbb{E}[S_i \pi_i (1-p_i)(1-T_i) Y_i].$$

Hence,  $\hat{\tau} \xrightarrow{P} \mathbb{E}[S_i \pi_i p_i T_i Y_i - S_i \pi_i (1-p_i)(1-T_i) Y_i]$ . We focus on the term on the right.

$$\begin{aligned} &\mathbb{E} \left\{ \pi_i \left( S_i p_i T_i Y_i - S_i (1-p_i)(1-T_i) Y_i \right) \right\} = \mathbb{E}_{\mathbf{W}} \left\{ \pi_i \mathbb{E} \left\{ S_i p_i T_i Y_i - S_i (1-p_i)(1-T_i) Y_i \mid \mathbf{W}_i \right\} \right\} \\ &= \mathbb{E} \left\{ \pi_i \Pr(S_i = 1 \mid \mathbf{W}_i) \mathbb{E} \left\{ p_i T_i Y_i - (1-p_i)(1-T_i) Y_i \mid S_i = 1, \mathbf{W}_i \right\} \right\} \\ &= \mathbb{E} \left\{ \pi_i \Pr(S_i = 1 \mid \mathbf{W}_i) \{ p_i \mathbb{E}[T_i Y_i \mid S_i = 1, \mathbf{W}_i] - (1-p_i) \mathbb{E}[(1-T_i) Y_i \mid S_i = 1, \mathbf{W}_i] \} \right\} \\ &= \mathbb{E} \left\{ \pi_i \Pr(S_i = 1 \mid \mathbf{W}_i) \left( \mathbb{E}[Y_i(1) \mid S_i = 1, \mathbf{W}_i] - \mathbb{E}[Y_i(0) \mid S_i = 1, \mathbf{W}_i] \right) \right\} \\ &= \mathbb{E} \left\{ \pi_i \Pr(S_i = 1 \mid \mathbf{W}_i) \mathbb{E}[Y_i(1) - Y_i(0) \mid S_i = 1, \mathbf{W}_i] \right\} = \mathbb{E} \left\{ \pi_i \Pr(S_i = 1 \mid \mathbf{W}_i) \mathbb{E}[Y_i(1) - Y_i(0) \mid S_i = 0, \mathbf{W}_i] \right\} \\ &= \mathbb{E} \left\{ \frac{\Pr(S_i = 0 \mid \mathbf{W}_i)}{\Pr(S_i = 0)} \mathbb{E}[Y_i(1) - Y_i(0) \mid S_i = 0, \mathbf{W}_i] \right\} \\ &= \int_{\mathbf{W}} \left\{ \frac{\Pr(S_i = 0 \mid \mathbf{W}_i)}{\Pr(S_i = 0)} \mathbb{E}[Y_i(1) - Y_i(0) \mid S_i = 0, \mathbf{W}_i] \right\} p(\mathbf{W}) d\mathbf{W} \\ &= \int_{\mathbf{W}} \mathbb{E}[Y_i(1) - Y_i(0) \mid S_i = 0, \mathbf{W}_i] p(\mathbf{W} \mid S_i = 0) d\mathbf{W} = \mathbb{E}[Y_i(1) - Y_i(0) \mid S_i = 0], \end{aligned}$$

where the first equality follows from the law of conditional expectation given  $\mathbf{W}$ , the second from the conditional expectation given  $S$ , the third from the linearity of expectation, the fourth

from the conditional expectation given  $T$ , the fifth from the linearity of expectation, the sixth from the definition of separating  $\mathbf{W}$ , the seventh from the definition of  $\pi$ , the eight from the rule of expectation, the ninth from Bayes rule, and the tenth from the rule of expectation.

## SI-3 Simulation Studies

We turn now to simulations to explore how well the proposed algorithm can recover the PATE. We first verify that our proposed algorithm can obtain a consistent estimator of the PATE. More importantly, we find that estimators based on estimated separating sets often have similar standard errors to the ones based on the true sampling set. Although our approach introduces an additional estimation step of finding separating sets to relax data requirements for the target population, it does not suffer from substantial efficiency loss. Both results hold with and without user constraints on what variables can be measured in the target population.

### SI-3.1 Simulation Design

In this subsection, we articulate our simulation design step by step. See the supplementary material for all the details on the simulation design.

**Pre-treatment Covariates and Potential Outcome Model.** To consider different types of separating sets, we assume the causal directed acyclic graph (DAG) in Figure SI-4 that encodes causal relationships among the outcome, the sampling indicator, and pre-treatment covariates. In this DAG, there are three conceptually distinct sets that we consider – (1) a sampling set,  $X4$  and  $X5$ , depicted in green, (2) a heterogeneity set,  $X2$  and  $X3$ , depicted in orange, and (3) the minimum separating set,  $X1$ , highlighted in purple. Three root nodes  $X1$ ,  $X6$ ,  $X7$  are normally distributed and other pre-treatment covariates are linear functions of their parents in the DAG. In particular, pre-treatment covariates are generated as follows.

$$\begin{aligned}
 X1 &\sim \mathcal{N}(0, 1) \\
 X2 &= 0.7 \times X1 + \sqrt{1 - 0.7^2} \times \epsilon_2 \\
 X3 &= 0.7 \times X1 + \sqrt{1 - 0.7^2} \times \epsilon_3 \\
 X4 &= 0.7 \times X1 + \sqrt{1 - 0.7^2} \times \epsilon_4 \\
 X5 &= 0.3 \times X9 + \sqrt{1 - 0.3^2} \times \epsilon_5 \\
 X6 &\sim \mathcal{N}(0, 1) \\
 X7 &\sim \mathcal{N}(0, 1) \\
 X8 &= -0.7 \times X2 + \sqrt{1 - 0.7^2} \times \epsilon_8 \\
 X9 &= 0.6 \times X1 + \sqrt{1 - 0.6^2} \times \epsilon_9
 \end{aligned}$$

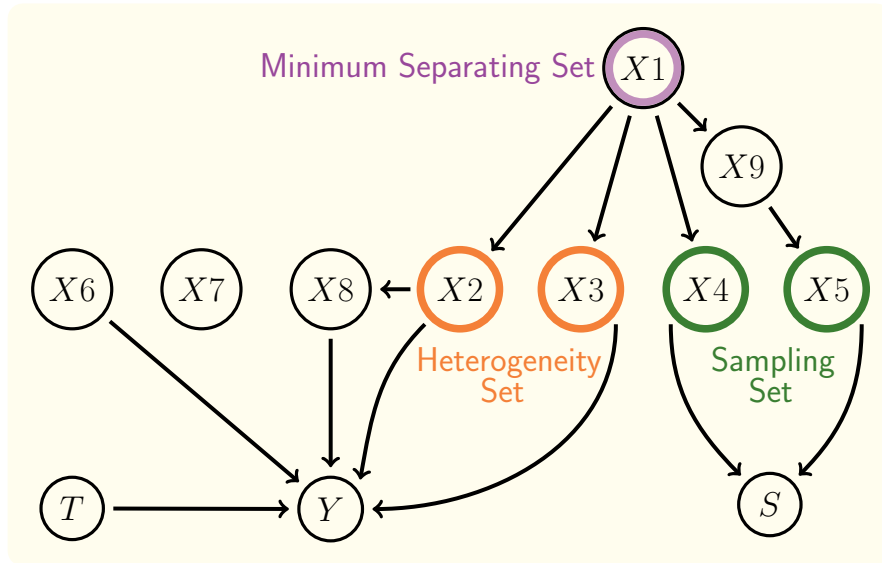


Figure SI-4: Causal DAG underlying the simulation study. Note: We consider three conceptually distinct sets (1) a sampling set,  $X_4$  and  $X_5$  (green), (2) a heterogeneity set,  $X_2$  and  $X_3$  (orange) and (3) the minimum separating set,  $X_1$  (purple). Three root nodes  $X_1$ ,  $X_6$ ,  $X_7$  are normally distributed and other pre-treatment covariates are linear functions of their parents.

where  $\epsilon_2, \epsilon_3, \epsilon_4, \epsilon_5, \epsilon_8, \epsilon_9$  are drawn independently and identically from a standard normal distribution,  $\mathcal{N}(0, 1)$ . This results in the following correlation structure for variables  $X_1 - X_9$ .

$$\text{cor}(\mathbf{X}) = \begin{pmatrix} 1.00 & -0.70 & 0.70 & 0.70 & -0.20 & 0.00 & 0.00 & 0.50 & -0.70 \\ -0.70 & 1.00 & -0.50 & -0.50 & 0.15 & 0.00 & 0.00 & -0.70 & 0.50 \\ 0.70 & -0.50 & 1.00 & 0.50 & -0.15 & 0.00 & 0.00 & 0.33 & -0.50 \\ 0.70 & -0.50 & 0.50 & 1.00 & -0.15 & 0.00 & 0.00 & 0.33 & -0.50 \\ -0.21 & 0.15 & -0.15 & -0.15 & 1.00 & 0.00 & 0.00 & -0.10 & 0.30 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 \\ 0.50 & -0.70 & 0.33 & 0.33 & -0.10 & 0.00 & 0.00 & 1.00 & -0.33 \\ -0.70 & 0.50 & -0.50 & -0.50 & 0.30 & 0.00 & 0.00 & -0.33 & 1.00 \end{pmatrix}$$

We then draw the potential outcomes as follows.

$$Y_i(T_i) = 5T_i + 10 \times X_{3i} \times T_i - 10 \times X_{2i} \times T_i + X_{6i} - 3 \times X_{8i} + \epsilon_i$$

where  $\epsilon_i \sim N(0, 1)$ . Thus, the true PATE is set to 5.

**Sampling Mechanism and Treatment Assignment.** We randomly sample a set of  $n$  units for a randomized experiment. The sampling mechanism is a logit model based on the sampling set,  $X_4$  and  $X_5$ . The treatment assignment mechanism is defined only for the

experimental sample ( $S_i = 1$ ). After being sampled into the experiment, every unit has the same probability of receiving the treatment  $\Pr(T_i = 1 \mid S_i = 1) = 0.5$ . For the sake of simplicity, we omit an arrow from the sampling indicator  $S$  to the treatment  $T$  in Figure 1.

In particular, we draw a sampling indicator  $S_i$  as follows. The second step scales the probability to be bounded away from zero and one.

$$\begin{aligned} S'_{i,lp} &= -20 \times X_{4i} + 20 \times X_{5i} \\ S_{i,lp} &= 0.25(S'_{i,lp} - \overline{S'_{lp}}) / sd(S'_{lp}) \\ S_i &= \frac{1}{1 + e^{-S_{i,lp}}} \end{aligned}$$

**Simulation Procedure.** We conduct 5000 simulations for each of six experimental sample sizes,  $n = \{100, 200, 500, 1000, 2000, 3000\}$ . Within each simulation, we first randomly sample  $n$  units for the experiment based on the sampling mechanism and randomly assign units to treatment according to the specified treatment assignment mechanism. We also randomly sample a target population of size  $m = 10000$ . We then estimate both an exact and a marginal separating set using the experimental data. An advantage of our method is that researchers can specify variables that cannot be measured in the target population. To illustrate this benefit, we also estimate a marginal separating set with a constraint that variable  $X1$  is unmeasurable in the target population, thus making the minimal separating set unobservable in the target population. We compare these sets to an oracle sampling set, oracle heterogeneity set, and oracle minimum separating set.

For each estimated and oracle set, we compute the PATE using the inverse probability weighting estimator described in Section 4.3. In the supplementary material, we repeat these simulations with a calibration estimator discussed in Hartman *et al.* (2015), and a linear regression projection estimator.

## SI-3.2 Results

We present results in Figure SI-5. Not shown in the graph are the results for the naive difference-in-means, which has significant bias ( $-1.0$ ). As expected, we see that the bias goes to zero for the oracle and estimated separating sets, and that the estimators are consistent for the PATE. More importantly, we see that estimators based the selected marginal separating sets (red), exact separating sets (dark blue), and marginal separating set with user constraints (pink) have similar standard errors to the oracle sampling set (green) and the oracle minimum separating set (purple). An estimator based on the oracle heterogeneity set (orange) has smaller standard errors than other estimators partly because it contains variables which are

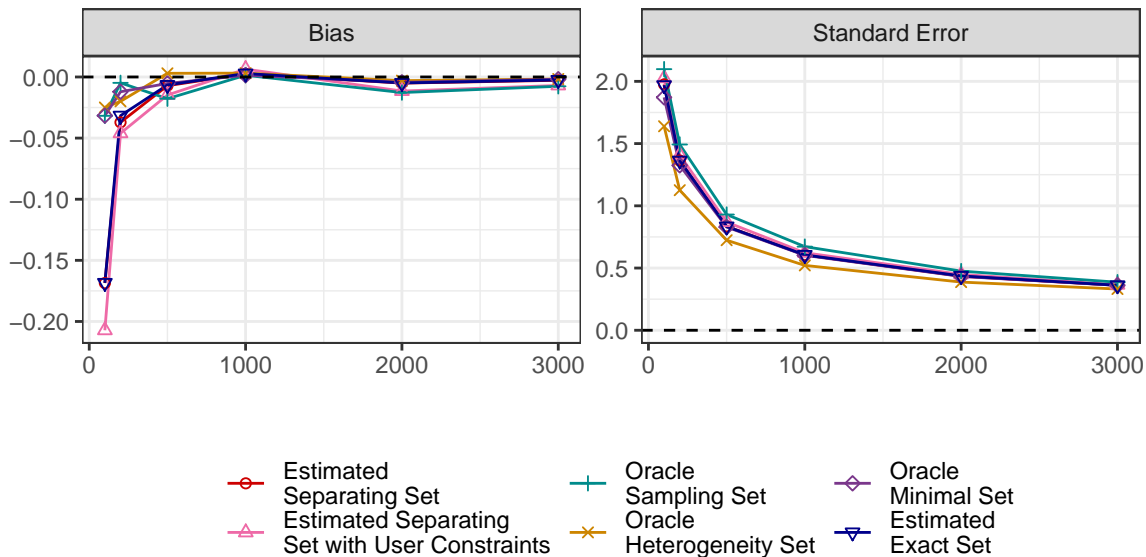


Figure SI-5: Simulation Results. Note: The left figure shows bias for the PATE and the right figure presents standard error estimates. As expected, bias is close to zero for all estimators. More importantly, estimators based the estimated separating sets (red) and estimated separating sets with user constraints (pink) have similar standard errors to the oracle sampling set (green) and the oracle minimum separating set (purple).

direct predictors of outcomes.

Figure SI-6 shows the breakdown of types of estimated separating sets. We group sets that are conceptually similar, and the frequency with which each set is chosen is presented. For example, if our algorithm selects the variables in the sampling set ( $X_4$  and  $X_5$ ) as well as an additional variable, we group these as “similar to” the sampling set. As can be seen, in these simulations as  $n$  gets large, over 75% of the time, the minimal separating set (purple) is selected. Small sample size can lead to the misestimation of the MRF, and therefore selection of inappropriate sets (gray) which do not remove bias — however, the rate at which inappropriate sets are selected drops off rapidly with sample size. In the supplementary material, we show that, when incorporating user constraints that make adjustment by the minimum separating set infeasible, the algorithm selects sets similar to the sampling and heterogeneity sets with higher frequency.

### SI-3.3 Additional Simulation Results

In the previous subsection, we discussed the breakdown of the different types of estimated separating sets in the simulated data generating process. Here we show the breakdown of types of estimated separating sets when incorporating user constraints in Figure SI-7. In this case,  $X_1$ , the alternative separating set, cannot be measured in the target population, we see that the algorithm selects the sampling and heterogeneity sets with higher frequency.

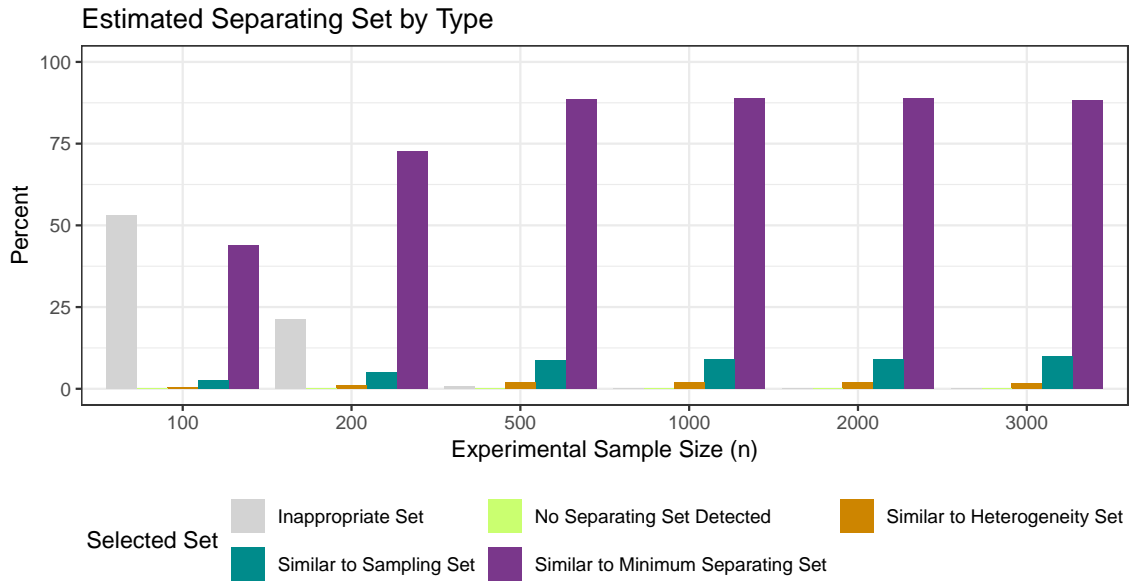


Figure SI-6: Types of Estimated Separating Sets. Note: We present the frequency of estimated separating sets by conceptual type. While the algorithm picks an inappropriate set when the sample size is small, as  $n$  increases, the most likely set is the minimal separating set.

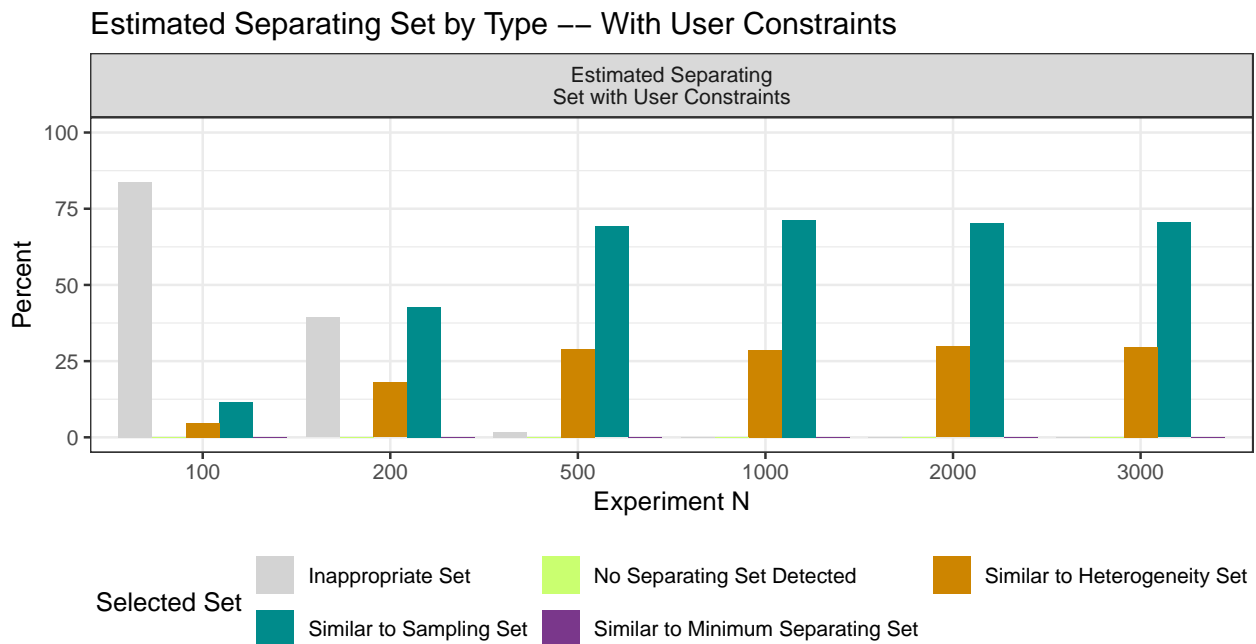


Figure SI-7: Type of Estimated Marginal Separating Set with User Constraints. Note: We present the frequency of estimated separating sets by conceptual type. With user constraints, the algorithm selects each of the other types of separating sets more frequently.

Figure SI-8 presents the bias and standard error result by selected estimated separating set type. We refer to sets that are “similar to” different conceptual sets in order to group sets that control for a specific type of separating sets, but which may include extra variables. For example, if the estimated set includes  $X_4$ ,  $X_5$ , and  $X_8$ , we say this is similar to a sampling set ( $X_4$  and  $X_5$ ). As theorems tell us, it doesn’t matter what type of separating sets the algorithm estimates in the experimental data, all of them produce unbiased estimates so long as the set is an appropriate separating set (see Figure SI-8). When an inappropriate set is chosen, which is common in the  $n = 100$  case but rare as  $n$  increases, we see that inappropriate sets do not reduce bias. As we expect, when estimated separating sets are similar to a heterogeneity set, standard errors are the smallest.

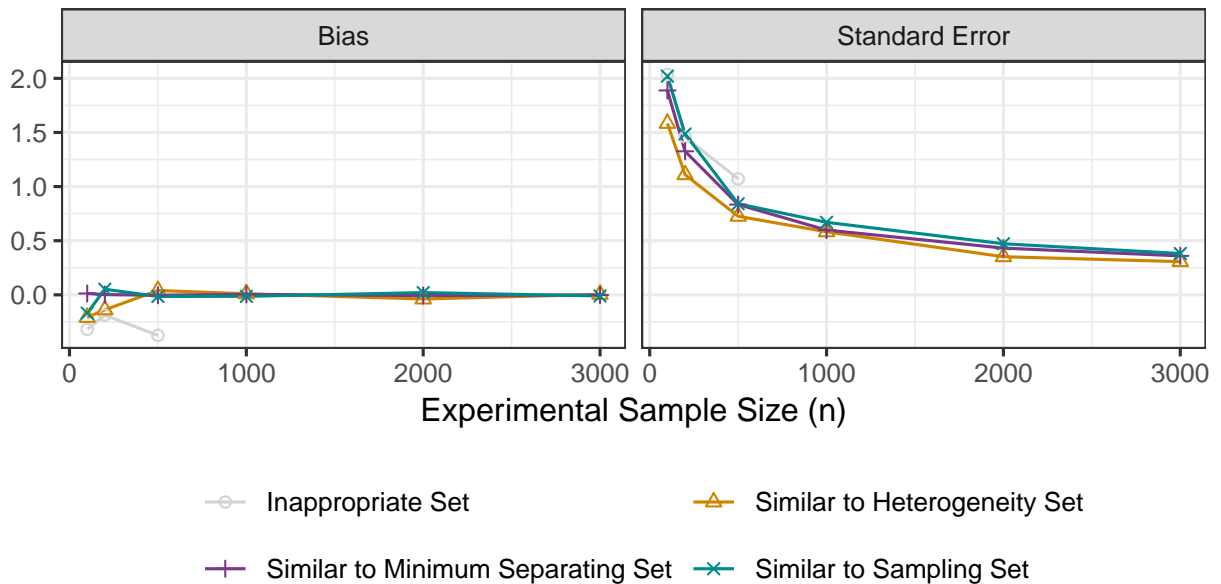


Figure SI-8: Simulation Results for Estimated Separating Set by Type. Note: The left figure shows bias for the PATE and the right figure presents standard error estimates. As expected, bias is close to zero for all estimators. Estimated sets are categorized by type: similar to oracle sampling set (green) and the oracle minimum separating set (purple) and oracle heterogeneity set (orange).

Finally, we present the simulation results for two alternative estimators in Figure SI-9, a calibration estimator and a linear regression projection. The calibration estimator matches population means for the estimated separating set using a maximum entropy (raking) algorithm (Hartman *et al.*, 2015). The linear projection estimator estimates a fully interacted linear regression model using the estimated separating set, and projects the model on the target population.



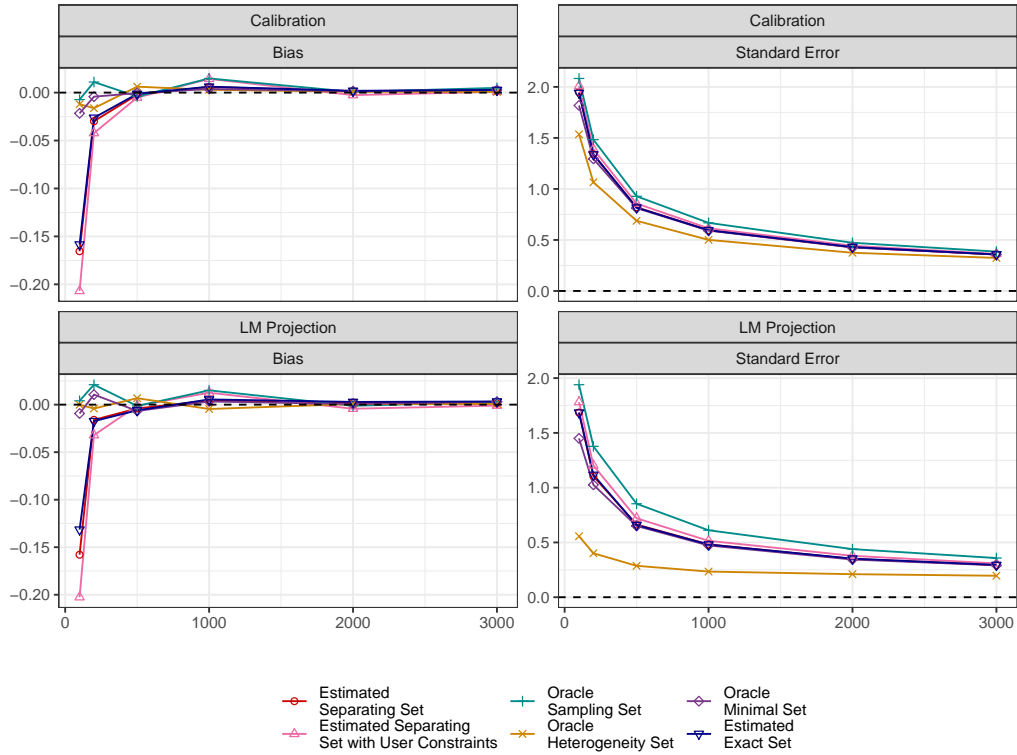


Figure SI-9: Simulation Results for Alternative Estimators. Note: The left figure shows bias for the PATE and the right figure presents standard error estimates. As expected, bias is close to zero for all estimators. More importantly, estimators based the estimated separating sets (red) and estimated separating set with user constraints (pink) have similar standard errors to the oracle sampling set (green) and the oracle minimum separating set (purple). An estimator based on the heterogeneity set (orange) has significantly smaller standard errors than other estimators, but this estimator might be unavailable in practice.